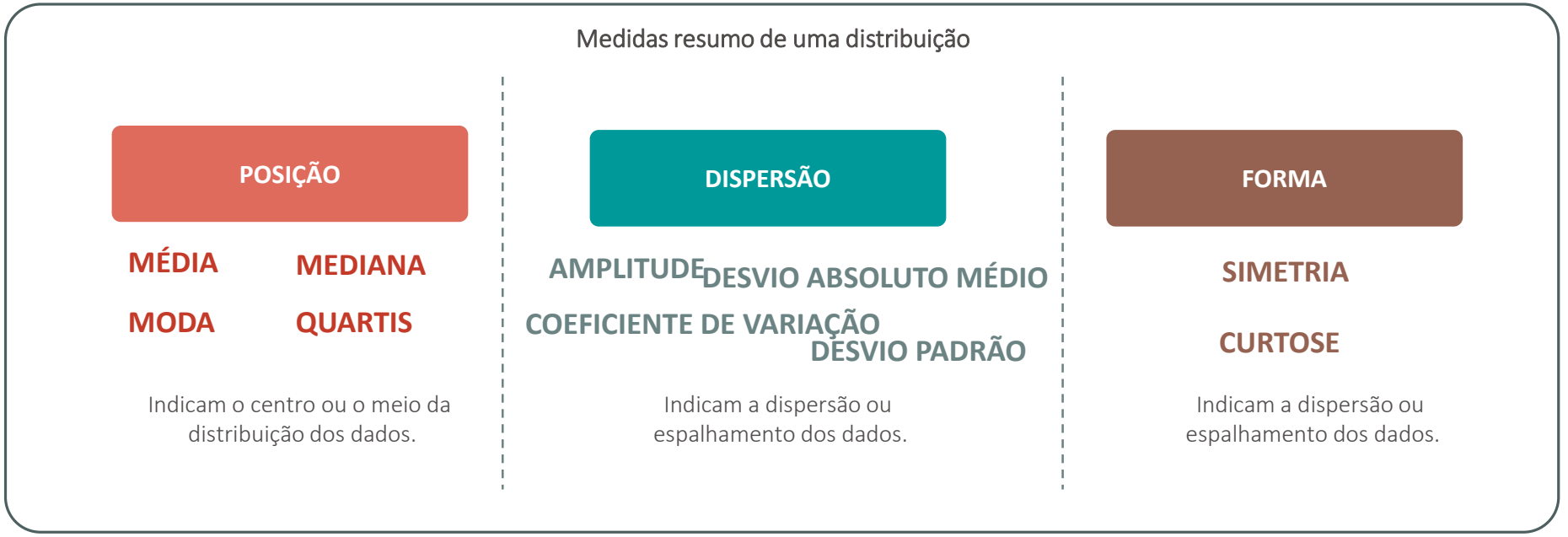
Estatística Básica I

MEDIDAS RESUMO

TUANY CASTRO

Conceito de distribuição

A distribuição representa o padrão de variação dos resultados que podem ser obtidos em um processo.



Medidas de Posição

As medidas de posição indicam o centro ou o meio da distribuição dos dados, ou seja, são medidas de tendência central.

MÉDIA

É a estatística mais utilizada para representar a locação dos dados. É a soma das observações dividida pelo número delas, considerada o ponto de equilíbrio de um conjunto de dados.

MEDIANA

É o termo central de uma sequência de valores colocados em ordem crescente.

MODA

É utilizada para designar o elemento de um conjunto de valores que aparece com maior frequência.

Exemplo

Suponha que parafusos a serem utilizados em tomadas elétricas são embalados em caixas rotuladas como contendo 100 unidades. Em uma construção, 15 caixas de um lote tiveram o número de parafusos contados, fornecendo os valores 98, 102, 100, 100, 99, 97, 96, 95, 99, 100, 100, 99, 100, 98, 100. Para essas caixas, quais os valores da média, mediana e moda?

MÉDIA

$$\bar{x} = \frac{98 + 102 + \cdots + 100}{15} = \frac{1483}{15}$$

$$\bar{x} = 98{,}9$$

MEDIANA

Dados ordenados:
95, 96, 97, 98, 98, 99, 99, 99, 100,100, 100, 100, 100, 100, 102

Número ímpar de observações: mediana é o 8^o termo

$$md(X) = 99$$

MODA

$$mo(X) = 100$$

Exemplo

Suponha que parafusos a serem utilizados em tomadas elétricas são embalados em caixas rotuladas como contendo 100 unidades. Em uma construção, 10 caixas de um lote tiveram o número de parafusos contados, fornecendo os valores 98, 102, 100, 100, 99, 97, 96, 95, 99, 100. Para essas caixas, quais os valores da média, mediana e moda?

MÉDIA

$$\bar{x} = \frac{98 + 102 + \cdots + 100}{10} = \frac{986}{10}$$

$$\bar{x} = 98{,}6$$

MEDIANA

Dados ordenados:
95, 96, 97, 98, 99, 99,
100,100, 100, 102

Número par de observações:
mediana entre o 5^o e o 6^o termo

$$md(X) = \frac{99 + 99}{2} = 99$$

MODA

$$mo(X) = 100$$

Calculando no R

Funções para cálculo no R:

- Média:

```
mean(dados$variável, na.rm=TRUE)
```

- Mediana:

```
median(dados$variável, na.rm=TRUE)
```

Calculando no R

- Moda:

Função para cálculo da moda:

```
mode <- function(x){
   ux <- unique(x)
   ux <- ux[which(is.na(ux)==FALSE)]
   ux[which.max(tabulate(match(x, ux)))]
}
```

Pedindo valor da moda:

```
mode(dados$variável)
```

Por que usar a mediana?

Suponha que, no exemplo dos parafusos, as observações das 10 caixas de um lote tiveram o número de parafusos contados iguais a 98, 102, 100, 100, 99, 97, 96, 95, 45, 100.

MÉDIA

$$\bar{x} = \frac{98 + 102 + \cdots + 100}{10} = \frac{932}{10}$$

$$\bar{x} = 93{,}2$$

MEDIANA

Dados ordenados:
45, 95, 96, 97, 98, 99,
100,100, 100, 102

Número par de observações:
mediana entre o 5^o e o 6^o termo

$$md(X) = \frac{98 + 99}{2} = 98{,}5$$

A mediana não é afetada por valores discrepantes (medida robusta).

Medidas de dispersão

Considere duas salas de aula em que foram observadas as seguintes notas:

Sala 1: 10, 9, 9, 8, 9

Sala 2: 9, 9, 9, 9, 9

Para as duas salas:

- Média: $\bar{x} = 9$
- Mediana: $md(X) = 9$
- Moda: $mo(X) = 9$

O que diferencia as duas salas? Dispersão dos dados

Medidas de dispersão

Amplitude

É definida como a diferença entre o maior e o menor valor do conjunto de dados.

No exemplo das salas:

Sala 1: 10, 9, 9, 8, 9
Amplitude: $\Delta = 10 - 8 = 2$

Sala 2: 9, 9, 9, 9, 9
Amplitude: $\Delta = 9 - 9 = 0$

Medidas de dispersão

DESVIO ABSOLUTO MÉDIO

Média dos desvios absolutos em relação à média.

VARIÂNCIA

Média dos desvios quadrados em relação à média.

DESVIO-PADRÃO

Raiz quadrada da variância.

Calculando no R

Funções para cálculo no R:

- Variância:

```
var(dados$variável, na.rm=TRUE)
```

- Desvio-padrão:

```
sd(dados$variável, na.rm=TRUE)
```

Medidas de dispersão

Coeficiente de variação

É a medida de variação dos dados em relação à média.

$$CV(X) = \frac{dp(X)}{\bar{X}}$$

Expressa a variabilidade tirando a influência da ordem de grandeza dos dados e permite a comparação da dispersão de distribuições diferentes.

Exercícios

1) Um estudante está procurando um estágio para o próximo ano. As companhias A e B têm programas de estágios e oferecem uma remuneração por 20 horas semanais com características (em salários mínimos) apresentadas abaixo. Qual a companhia mais adequada?

Companhia	A	B
média	2,5	2
mediana	1,7	1,9
moda	1,5	1,9

2) Você está indeciso em comprar um notebook e decide avaliar algumas informações estatísticas, fornecidas pelo fabricante, sobre a duração da bateria (em horas). Com que marca você ficaria?

Marca	GA	FB	HW
Média	8,0	8,2	8,0
Mediana	8,0	9,0	7,0
Desvio-Padrão	1,0	1,5	2,5

Exercícios

3) Num experimento, dois grupos de15 mulheres foram submetidas a dois diferentes tipos de dietas para emagrecimento. Os dados referentes à perda de peso se encontram em 'dados_dieta.csv'. Qual seria a melhor dieta?

4) O arquivo *cancer.csv* contém os dados de uma pesquisa sobre incidência de câncer e é apresentado em 4 colunas representando as seguintes variáveis de interesse:

- coluna 1: identificação do paciente
- coluna 2: diagnóstico:
  - 1 = falso-negativo
  - 2 = negativo
  - 3 = positivo
  - 4 = falso-positivo

Exercícios

- coluna 3: idade
- coluna 4: glicose (GL)

A) Obtenha as medidas de posição e de dispersão para as variáveis Idade e Glicose.

B) Repita o item anterior para cada tipo de diagnóstico.

5) Os dados do arquivo *comunidade.csv* contém parte dos dados de uma pesquisa sobre aspectos socioeconômicos e culturais de comunidades de baixa renda da região do Butantã em São Paulo. Os dados estão organizados da seguinte forma:

- Coluna 1: número do questionário (Num)
- Coluna 2: Sexo:
  - 1: masculino
  - 2: feminino

Exercícios

Coluna 3: Idade que começou a trabalhar, em anos

- Colune 4: Série em que parou de estudar
  - N: Não parou de estudar
  - F: parou de estudar no ensino fundamental
  - M: parou de estudar no ensino médio

A) Obtenha as medidas de posição e dispersão para a variável Idade.

B) Repita o item A para cada uma das categorias da série em que parou de estudar. Existem diferenças entre as categorias?

C) Baseado nas variáveis Sexo e Idade, você diria que os homens começam a trabalhar mais cedo?

Exercícios

6) Para avaliar o funcionamento das máquinas de uma empresa, foram coletados durante 332 dias uma amostra de 200 peças para avaliar o total de peças defeituosas.

- **A)** Calcule a média e o desvio-padrão do número de peças defeituosas.
- **B)** Para efeito de análise, decidiu-se desprezar os valores que se distanciassem da média amostral por mais de dois desvios-padrão, isto é, só serão considerados os valores no intervalo $\bar{x} \mp 2dp(x)$. Recalcule o item (A) e comente.